RL hw1

1.(a)

$$\pi_\theta(\cdot\,|s) = \frac{exp(\theta_\cdot)}{exp(\theta_a) + exp(\theta_b) + exp(\theta_c)}$$

$$\frac{\partial log\pi_\theta(a|s)}{\partial\theta_a} = 1 - \frac{exp(\theta_a)}{exp(\theta_a) + exp(\theta_b) + exp(\theta_c)} = 1 - \pi_\theta(a|s)$$

$$\frac{\partial log\pi_\theta(a|s)}{\partial\theta_b} = -\frac{exp(\theta_b)}{exp(\theta_a) + exp(\theta_b) + exp(\theta_c)} = -\pi_\theta(b|s)$$

$$\frac{\partial log\pi_\theta(a|s)}{\partial\theta_c} = -\frac{exp(\theta_c)}{exp(\theta_a) + exp(\theta_b) + exp(\theta_c)} = -\pi_\theta(c|s)$$

And so on, we can get $\frac{\partial log\pi_\theta(b|s)}{\partial\theta_\cdot}$ and $\frac{\partial log\pi_\theta(c|s)}{\partial\theta_\cdot}$ by similar method.

$$\pi_\theta(a|s) = \frac{exp(\theta_a)}{exp(\theta_a) + exp(\theta_b) + exp(\theta_c)} = 0.1, \pi_\theta(b|s) = 0.5, \pi_\theta(c|s) = 0.4$$

$$\widehat{\nabla}V = r(s,\cdot)\nabla_\theta\pi_\theta(\cdot\,|s)$$

$$\widehat{\nabla}V_a = 100\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix}, \widehat{\nabla}V_b = 98\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix}, \widehat{\nabla}V_c = 95\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix}$$

$$\mathbb{E}\big[\widehat{\nabla}V\big] = 0.1\widehat{\nabla}V_a + 0.5\widehat{\nabla}V_b + 0.4\widehat{\nabla}V_c = \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix} \blacksquare$$

$$\mathbb{E}\left[\left(\widehat{\nabla}V - \mathbb{E}\big[\widehat{\nabla}V\big]\right)\left(\widehat{\nabla}V - \mathbb{E}\big[\widehat{\nabla}V\big]\right)^T\right] = \mathbb{E}\left[\widehat{\nabla}V\left(\widehat{\nabla}V\right)^T\right] - \mathbb{E}\big[\widehat{\nabla}V\big]\mathbb{E}\big[\widehat{\nabla}V\big]^T$$

$$= 0.1 \times 100^2\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix}\begin{bmatrix}0.9\\-0.5\\-0.4\end{bmatrix}^T + 0.5 \times 98^2\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix}\begin{bmatrix}-0.1\\0.5\\-0.4\end{bmatrix}^T$$

$$+ 0.4 \times 95^2\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix}\begin{bmatrix}-0.1\\-0.5\\0.6\end{bmatrix}^T - \begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}\begin{bmatrix}0.3\\0.5\\-0.8\end{bmatrix}^T$$

$$= \begin{bmatrix} 810 & -450 & -360 \\ -450 & 250 & 200 \\ -360 & 200 & 160 \end{bmatrix} + \begin{bmatrix} 48.02 & -240.1 & 192.08 \\ -240.1 & 1200.5 & -960.4 \\ 192.08 & -960.4 & 968.32 \end{bmatrix}$$

$$+ \begin{bmatrix} 36.1 & 180.5 & -216.6 \\ 180.5 & 902.5 & -1083 \\ -216.6 & -1083 & 1299.6 \end{bmatrix} - \begin{bmatrix} 0.09 & 0.15 & -0.24 \\ 0.15 & 0.25 & -0.4 \\ -0.24 & -0.4 & 0.64 \end{bmatrix}$$

$$= \begin{bmatrix} 894.03 & -509.75 & -384.28 \\ -509.75 & 2352.75 & -1843 \\ -384.28 & -1843 & 2227.28 \end{bmatrix} \blacksquare$$

1.(b)

$$V^{\pi_\theta}(s) = 0.1 \times 100 + 0.5 \times 98 + 0.4 \times 95 = 97$$

$$\hat{\nabla} V_a = (100 - 97) \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}, \hat{\nabla} V_b = (98 - 97) \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}, \hat{\nabla} V_c = (95 - 97) \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}$$

$$\mathbb{E}[\hat{\nabla} V] = 0.1\hat{\nabla} V_a + 0.5\hat{\nabla} V_b + 0.4\hat{\nabla} V_c = \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \blacksquare$$

$$\mathbb{E}\left[\left(\hat{\nabla} V - \mathbb{E}[\hat{\nabla} V]\right)\left(\hat{\nabla} V - \mathbb{E}[\hat{\nabla} V]\right)^T\right] = \mathbb{E}\left[\hat{\nabla} V\left(\hat{\nabla} V\right)^T\right] - \mathbb{E}[\hat{\nabla} V]\mathbb{E}[\hat{\nabla} V]^T$$

$$= 0.1 \times 3^2 \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix} \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}^T + 0.5 \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix} \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}^T$$

$$+ 0.4 \times (-2)^2 \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix} \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}^T - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}^T$$

$$= \begin{bmatrix} 0.729 & -0.405 & -0.324 \\ -0.405 & 0.225 & 0.18 \\ -0.324 & 0.18 & 0.144 \end{bmatrix} + \begin{bmatrix} 0.005 & -0.025 & 0.02 \\ -0.025 & 0.125 & -0.1 \\ 0.02 & -0.1 & 0.08 \end{bmatrix}$$

$$+ \begin{bmatrix} 0.016 & 0.08 & -0.096 \\ 0.08 & 0.4 & -0.48 \\ -0.096 & -0.48 & 0.576 \end{bmatrix} - \begin{bmatrix} 0.09 & 0.15 & -0.24 \\ 0.15 & 0.25 & -0.4 \\ -0.24 & -0.4 & 0.64 \end{bmatrix}$$

$$= \begin{bmatrix} 0.66 & -0.5 & -0.16 \\ -0.5 & 0.5 & 0 \\ -0.16 & 0 & 0.16 \end{bmatrix} \blacksquare$$

1.(c)

$$\mathbb{E}\left[\left(\hat{\nabla}V_B - \mathbb{E}[\hat{\nabla}V_B]\right)\left(\hat{\nabla}V_B - \mathbb{E}[\hat{\nabla}V_B]\right)^T\right] = \mathbb{E}\left[\hat{\nabla}V_B\left(\hat{\nabla}V_B\right)^T\right] - \mathbb{E}[\hat{\nabla}V_B]\mathbb{E}[\hat{\nabla}V_B]^T$$

$$= 0.1 \times \left(100 - B(s)\right)^2 \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}\begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}^T$$

$$+ 0.5 \times \left(98 - B(s)\right)^2 \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}\begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}^T$$

$$+ 0.4 \times \left(95 - B(s)\right)^2 \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}\begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}^T - \begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}\begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}^T$$

$$tr\left(\mathbb{E}\left[\left(\hat{\nabla}V_B - \mathbb{E}[\hat{\nabla}V_B]\right)\left(\hat{\nabla}V_B - \mathbb{E}[\hat{\nabla}V_B]\right)^T\right]\right)$$

$$= tr\left(0.1 \times \left(100 - B(s)\right)^2 \begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}\begin{bmatrix} 0.9 \\ -0.5 \\ -0.4 \end{bmatrix}^T\right)$$

$$+ tr\left(0.5 \times \left(98 - B(s)\right)^2 \begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}\begin{bmatrix} -0.1 \\ 0.5 \\ -0.4 \end{bmatrix}^T\right)$$

$$+ tr\left(0.4 \times \left(95 - B(s)\right)^2 \begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}\begin{bmatrix} -0.1 \\ -0.5 \\ 0.6 \end{bmatrix}^T\right)$$

$$- tr\left(\begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}\begin{bmatrix} 0.3 \\ 0.5 \\ -0.8 \end{bmatrix}^T\right)$$

$$= 0.1 \times \left(100 - B(s)\right)^2 (0.81 + 0.25 + 0.16)$$

$$+ 0.5 \times \left(98 - B(s)\right)^2 (0.01 + 0.25 + 0.16)$$

$$+ 0.4 \times \left(95 - B(s)\right)^2 (0.01 + 0.25 + 0.36)$$

$$- (0.09 + 0.25 + 0.64)$$

$$= 0.122\left(100 - B(s)\right)^2 + 0.21\left(98 - B(s)\right)^2 + 0.248\left(95 - B(s)\right)^2 - 0.98$$

$$= 0.58B(s)^2 - 112.68B(s) + 5474.06$$

$$minimum\ when\ B(s) = \frac{-b}{2a} = \frac{112.68}{2 \times 0.58} = \frac{112.68}{1.16} = \frac{2817}{29} \blacksquare$$

2.

$$\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[f(\mathrm{s},a)] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\left[\sum_a \pi_\theta(a|s)f(s,a)\right]$$

$$= \frac{1}{1-\gamma}\left[\sum_s d_\mu^{\pi_\theta}(s)\sum_a \pi_\theta(a|s)f(\mathrm{s},a)\right]$$

$$= \frac{1}{1-\gamma}\sum_s \mathbb{E}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi_\theta)\sum_a \pi_\theta(a|s)f(\mathrm{s},a)\right]$$

$$= \mathbb{E}_{s_0\sim\mu}\left[\sum_s\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi_\theta)\sum_a \pi_\theta(a|s)f(\mathrm{s},a)\right]$$

$$= \sum_\tau \mu(s_0)\sum_{t=0}^{\infty}\sum_s \gamma^t P(s_t = s|s_0,\pi_\theta)\sum_a \pi_\theta(a|s)f(\mathrm{s},a)$$

$$= \sum_\tau \mu(s_0)\sum_{t=0}^{\infty}\gamma^t f(\mathrm{s}_t,a_t)[\pi_\theta(a_0|s_0)P(s_1|s_0,a_0)\pi_\theta(a_1|s_1)\cdots]$$

$$= \mathbb{E}_{\tau\sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t f(\mathrm{s}_t,a_t)\right] \blacksquare$$

3.1

$$V(S) = P_S\big(R_S + V(S)\big) + P_T R_T$$
$$(1-P_S)V(S) = P_S R_S + P_T R_T$$
$$V(S) = \frac{P_S R_S + P_T R_T}{(1-P_S)} = \frac{P_S R_S + P_T R_T}{P_T} = \frac{P_S}{P_T}R_S + R_T \blacksquare$$

3.2

$$\mathbb{E}_\tau\big[\hat{V}_{MC}(S;\tau)\big] = \sum_{k=0}^{\infty} P_T {P_S}^k\left(\frac{R_S + 2R_S + 3R_S + \cdots + kR_S + (k+1)R_T}{k+1}\right)$$

$$= \sum_{k=0}^{\infty} P_T {P_S}^k\left(\frac{k(k+1)}{2}\frac{R_S}{k+1} + R_T\right)$$

$$= \sum_{k=0}^{\infty} P_T {P_S}^k\left(\frac{k}{2}R_S + R_T\right)$$

$$= P_T \left[\frac{R_S}{2} \sum_{k=0}^{\infty} k {P_S}^k + R_T \sum_{k=0}^{\infty} {P_S}^k \right]$$

$$= P_T \left[\frac{R_S}{2} \frac{P_S}{(1 - P_S)^2} + R_T \frac{1}{1 - P_S} \right]$$

$$= P_T \left[\frac{R_S}{2} \frac{P_S}{(P_T)^2} + R_T \frac{1}{P_T} \right]$$

$$= \frac{P_S}{2P_T} R_S + R_T \blacksquare$$